TO BE REVISED (FEBRUARY 2016)

Subspace transmission:

Desynchronized speech-gesture signals still get the message across

Carolin Kirchhof

Faculty of Linguistics and Literary Studies, Bielefeld University

Universitaetsstr. 25, 33613 Bielefeld, Germany

ckirchhof@uni-bielefeld.de

Abstract

Two sets of studies researched the perceptual integration of desynchronized speech and gesture in listeners. In the first set, participants rated how natural asynchronies in multimodal stimuli with varying face obscurity appeared to them, both with speech before and after the gesture for up to 600 ms. In the second set, participants were asked to re-synchronize speech and gesture with a slider. Both sets of studies show that the synchrony of the two modalities is far less significant in perception than was assumed a priori through the observation of production. Speech may precede or follow gesture by ±500 ms or more and listeners might not even notice. Strict speech-gesture synchrony appears to be a mere production phenomenon.

Key words: perception, audiovisual integration, AVI, speech-gesture synchrony

About the author

Carolin Kirchhof, M.A., is a researcher and lecturer for linguistics at the Faculty of Linguistics and Literary Studies at Bielefeld University. Her PhD project with J.P. de Ruiter is concerned with the perception of speech-gesture synchrony in natural dyadic communication and with the conceptual affiliation of speech and gesture. Having graduated in British and American Studies and Text Technology, she began concentrating on gesture after participating in the 2009 LSA Summer Institute at Berkeley and research stays with David McNeill and Susan Duncan at the Psychology department of the University of Chicago.

Index of Figures

1. Introduction

Spontaneous gestures and the associated speech are produced approximately simultaneously (e.g. Kendon, 1980, 2004; McNeill, 1985, 2005) and numerous studies have engaged in analyzing the significance of synchronized production for meaning creation: There is a semantic connection between the two modalities (e.g. Kirchhof, 2011; de Ruiter, 2000; Krauss, 2000; de Ruiter & Wilkins, 1998; Schegloff, 1984). Presumably, the bimodal synchrony in production is deemed highly relevant for perception, since it is being programmed into virtual agents, robots, etc. (e.g. Wheatland, Wang, Song, Neff, Zordan, & Jörg 2015; Kopp & Wachsmuth 2004; Bergmann & Kopp 2009). The focus in gesture research, strikingly in contrast to other areas of psycholinguistics, has mainly been on production rather than on perception (e.g. Feyereisen, 2007; McNeill, 2005, 1985; Kendon, 2004, 1980). This paper will contribute to the research on speech-gesture perception by studying how listeners perceive asynchronies in naturally co-produced speech and gestures by eliciting their preferences using different methodologies.

Several studies have looked at the comprehension of speech and gestures (e.g. Holler, Shovelton, & Beattie, 2009; Gullberg, & Kita, 2009; Gullberg & Holmqvist, 2006; Alibali, Heath, & Myers, 2001). The effect of relative timing on comprehension has only recently been addressed explicitly (Habets, Kita, Shao, Özyürek, & Hagoort, 2011; Özyürek, Willems, Kita, & Hagoort, 2007). While strong asynchrony between the modalities during speaking, registered through self-monitoring by the speaker, prompts them to repair their utterance (e.g. Seyfeddinipur & Kita, 2010), the listener is expected to disregard or internally align the smaller asynchronies to ensure proper comprehension. Several studies in the area of psychophysics (e.g. Nishida, 2006; Fujisaki & Nishida, 2005) have found a time window for the perceptual alignment of visual and auditory signals, the so-called audiovisual integration (AVI) window. Studies such as those on speech-lip asynchrony by McGurk and MacDonald (1976) and Massaro, Cohen, & Smeele (1996; also Vatakis, Navarra, Soto-Faraco, & Spence, 2008; Massaro & Cohen, 1993) have inspired research in the gesture field, for instance by Habets et al. (2011), but no studies have yet been conducted with natural data and no dyadic situations have been analyzed. The two sets of studies presented in this paper will hopefully set a starting

point for further investigations into the perception of asynchronies of speech and gesture. As there is a semantic bond between the two modalities, a certain window of audiovisual integration (AVI) should be expected. The studies will investigate what listeners perceive as well as their ability to reproduce what they assume happens in production: If they cannot tell how speech and gesture should be synchronized in reality, asking them what window of bimodal synchrony they prefer seems futile.

1.1. Terminology

The studies presented in this paper contrast semiotic signals in the form of speech-gesture utterances with non-semiotic signals. The word ‘gesture’ will be used for spontaneous movements of the hands co-occuring with spontaneous natural speech. The terminology for the different gesture types used is based on the semantic categorizations by McNeill (1992), not strictly adhering to it in every aspect. The following categories of signals are distinguished:

Iconic gestures: Gestures related to the speech they accompany in shape and other physical properties highlighting certain semiotic aspects of this speech. This includes gestures with metaphoric function (cf. McNeill, 1992; also de Ruiter 2000; p. 285), but not gestures used for facilitating speech or gestures used in turn management.

Emblematic gestures: Gestures that do not need speech to disambiguate their meaning and are, hence, standalone lexical items.

Deictic gestures: Pointing gestures; regardless of hand configuration.

Cause-and-effect signals: In the context of non-speech, non-semiotic audiovisual stimuli, this term describes sounds directly caused by a motion or mechanism, such as the sound emerging from clapping hands or from knocking on wood. This categorization does not include speech-only utterances, but is explicitly distinguished from these.

1.2. Previous research on the perception of audio-visual signals

Speech-lip (a)synchronies

We perceive a causal connection between the events and sounds of clapping, of thunder and lightning, or of ringing a bell. We feel the same about mouth movements and speech and are irritated by low-quality dubbing or audio-video delays, yet research on the

perceived synchrony of speech and gesture has only recently garnered more attention, the perception of audiovisual synchrony in general, and on speech-lip synchrony in particular, has already been a topic in psychophysics and phonetics for nearly half a century. McGurk and MacDonald (1976), for instance, described how subjects presented with two different CV-syllables (e.g. /ga/ & /ba/) simultaneously via the audio and video channels perceived the syllables as fused percepts (e.g. /da/). This finding demonstrates that the sounds of speech are not the only factors for the listener in communication, and that audiovisual synchrony plays a major role. This has also been established by Fujisaki and Nishida (2005), among others, for cause-and-effect stimuli of physical events such as light and beep signals.

Based on the so-called 'McGurk effect', Massaro and Cohen (1993) tested the perception of CV-clusters and vowels at different asynchronies of up to ±200 ms. The temporal range in which the bimodal stimuli were fused by the subjects can be considered the window of AVI. In order to further specify this window, Massaro, Cohen, & Smeele (1996) conducted experiments with varying and slightly larger asynchronies. Next to two identification tasks, in which subjects had to tell whether stimuli were in synchrony, they also used a fuzzy-logical model of perception (FLMP) which assumed the video and audio to be synchronous. The model "predict[ed] integration across different asynchronies as long as the two modalities [were] perceived as belonging to the same perceptual event", i.e. to one stimulus (p. 1778, see also Fujisaki and Nishida, 2005; van Wassenhove, Grant, & Poeppel, 2002). Massaro et al. (1996) used synthetic speech stimuli in addition to those from natural language. A polygon facial model displayed the randomized stimuli of modified McGurk-pairs to the participants. The tested asynchronies were varied in seven steps within ± 267 ms and additionally at ±533 ms. The authors concluded that an AVI breakdown would occur at asynchronies of about ± 500 ms while integration would be optimal within a window of ± 200 ms.

Van Wassenhove, Grant, & Poeppel (2007) used stimuli with fourteen steps of 33 ms in which the audio onset was put before or after the onset of the video up to discrepancies of ± 467 ms. These stimuli were used in an *identification task* as well as in a *simultaneity judgment task*, both of which were completed in succession by each participant. As in the original experiment by McGurk and MacDonald (1976), only natural speech and video recordings were

used. The participants in Van Wassenhove, Grant, & Poeppel (2007) chose between three possible percepts in a multiple-choice fashion in the identification task: the actual audio signal (/ga/ above), the actual video signal (/ba/ above) or the 'fused McGurk percept' (/da/ above). In the simultaneity judgment task, participants were then asked to determine whether audio and video were in synchrony. They had to choose between "simultaneous" and "successive", regardless of order. The window of AVI as judged from the responses with the fused percept, i.e. what the listener perceives from the audiovisual stimulus, reached from asynchronies of the audio 67 ms before the visual to 267 ms of the audio after the visual (range 334 ms). The subjects accepted a smaller AVI window (-73 ms to +131 ms) for the stimuli in which the audio and video contained identical syllables. Van Wassenhove et al. (2007) deduced a "maximal true bimodal fusions cluster within ~200 ms" (p. 604) from this. The identification task gave results well below the estimated breakdown of AVI at asynchronies of more than 500 ms (cf. Massaro et al. 1996). Van Wassenhove et al. (2007) conclusively accepted a window of about 200 ms[1] for general alignment but assumed that "to allow the extraction of modality-specific information", tighter synchrony was necessary (p. 605).

The question arises whether the findings on the temporal limits of the AVI of lip-speech signals described in Massaro and Cohen (1993), Massaro et al. (1996), and van Wassenhove et al. (2007), among others, also apply to the AVI of gesture and speech. It has been established that the subjectively perceived simultaneity varies across levels of asynchrony and that the circumstances under which one is confronted with stimuli, e.g. in an experimental setting or in real life, are relevant to integration (see also Fujisaki & Nishida, 2005; Nishida, 2006). Delays as well as advances of the audio or video channels are integrable by the listener. The visual and auditory modalities are integrated into a fused percept between an audio advance of 30 ms and an audio delay of 170 ms (van Wassenhove et al., 2007). A general AVI of bimodal syllables is possible at asynchronies of ±150 to ±250 ms while a significant breakdown in the perceptual alignment might be expected between ±250 ms and ±500 ms (Massaro et al., 1996).

[1] No exact numbers given by authors. Possibly ranging from −73 ms to +131 ms.

Speech-gesture (a)synchronies

While most research in the field of gesture focuses on *production*, within about the last fifteen years there has been an increase of studies on the *perception* of multimodal speech, e.g. by Gullberg & Holmqvist (2006) and Alibali et al. (2001). Their focus has mainly been on proving that listeners are capable of information uptake from gestures, for instance by showing pictures, cartoons, or even gesture clips before or with speech stimuli to listeners and then questioning them about these. Neuroscientific methods to research AVI as they have been applied to audiovisual speech perception (e.g. Ojanen, 2005; Callan, Jones, Munhall, Kroos, Callan, & Vatikiotis-Bateson, 2004; Bushara, Grafman, & Hallett, 2001) have also been a recent development in gesture research (e.g. Özyürek et al., 2007; Habets et al., 2011).

Özyürek et al. (2007) monitored participants for ERP while having them watch videos of spoken sentences including gestures. Their methodology followed Holler et al. (2009: the stimuli showed an actor performing previously observed iconic gestures (cf. example in Table 1: Example of stimulus construct used by Habets et al. (2011, p. 1849).). The gestures were manually synchronized at the stroke with complementing or conflicting verbs within selected sentences “because in 90% of natural speech-gesture pairs the stroke coincide[s] with the relevant speech segment” (Özyürek et al. 2007, p. 610, after McNeill 1992). The preceding part of the sentence served as the prime and the paired prosodic peak (e.g. pitch accent) and gesture stroke as the target for the ERP. At the point of simultaneous exposure, the listeners showed about the same ERP-response to all target stimuli: “[i]n all conditions, the N400 component reached its peak around 480 msec” (p. 612), with or without semantic congruency. The researchers interpreted these homogeneous results as hinting at a non-sequential AVI of speech and gesture: The integration might happen in parallel, as has been found in speech-lip research (p. 613). The findings by Özyürek et al. (2007) are highly relevant for researching the AVI of speech and gesture. As with other studies presented here, the stimuli, which were recorded using actors, were of non-natural and deliberate speech and gestures. Even artificially incongruent speech and gestures were integrated as if they were congruent. This agrees, for instance, with the findings by Cassell et al. (1999), who deducted that gestures are not only registered by the listener but that even ‘mismatched’ information is taken from them.

Habets et al. (2011) followed up on Özyürek et al. (2007) and added audio offsets to the experiments and expanded on the matter of semantic congruency. Their stimuli were created by combining separately recorded audio and video clips of verbs congruent and incongruent with the paired gestures (see Table 1: Example of stimulus construct used by Habets et al. (2011, p. 1849).); the channels were synchronized at the prosodic peak and gesture stroke or the audio was delayed after the video.

Table 1: Example of stimulus construct used by Habets et al. (2011, p. 1849).

	Target Words	
Target Gesture	*Match*	*Mismatch*
(1) The two fists are placed on top of each other, as if to hold a club, and they move away from the body twice.	Battering	Hurdling

Across brain regions, the stimuli produced similar results in the participants for the synchronized condition as for when the audio was delayed by 160 ms. The semantic mismatches triggered significantly higher activity, i.e. more complex integration ($p < .05$; Habets et al. 2011; p. 1851). The authors concluded from the lack of an N400 effect at an audio delay of 360 ms that "gesture interpretation might not be influenced by the information carried by speech" (p. 1852). They also claimed that "speech and iconic gestures are most effectively integrated when they are fairly precisely coordinated in time" (p. 1853). For combinations of single words and gestures that do not naturally co-occur, the study by Habets et al. (2011) supported the findings by Özyürek et al. (2007) on incongruent speech-gesture signals. The ERP results did not testify to what happens in complete, naturally co-occurring speech-gesture utterances, and the AVI window for single words with gestures might extend to somewhere between an auditory delay of 160 ms and 360 ms. It is also not quite clear from Habets et al. (2011) what happens to AVI when the speech precedes the gesture, but the authors deduce that "the interpretation of the gesture was fixed before the speech onset" in their study (p. 1852).

Özyürek et al. (2007) showed semantic congruency was not a factor when the modalities are synchronized at prosodic peak and gesture stroke onset, even when a contextual sentence preceded the critical stimulus. This is compatible with van Wassenhove et al. (2007), who found only a minimal difference of about 30 ms between congruent and incongruent signals

at which an audio advance was integrated. Habets et al. (2011) investigated “the aspect of semantic integration of gesture and speech” (p.1846). Since they used artificial speech-gesture pairs (p.1848), their results can only hint at the integration of naturally co-produced utterances. Also, as in Özyürek et al. (2007), the forced synchrony of the modalities was helpful for an ERP analysis but could have made the stimuli seem even more unnatural. The cutting off of the preparation phase of the gestures could also have influenced their results.

1.3. Impact of Previous Research on Methodology

Following up on the phenomenon of the McGurk effect, research has provided several approximations to the possible and optimal windows of AVI for speech-lip stimuli. Among others, van Wassenhove et al. (2007) found fused percepts between an audio advance of 30 ms and an audio delay of 170 ms from the lip movement. Massaro et al. (1996) expanded this window of audiovisual syllable integration to ± 267 ms, with a breakdown of integration occurring beyond this window. In contrast to these studies on speech-lip perception, Habets et al. (2011) as well as Özyürek et al. (2007) excluded the possibility of an audio advance before the gesture. Contributing to specifying the range of AVI for speech-gesture signals, their results do show that gestures coming up to 160 ms before the speech are integrated by the listener, but that at an advance of 360 ms, the semantic integration process differs, as was revealed by the N400 ERP component.

There are several methodological shortcomings in the research discussed better to be avoided when studying natural communication. Firstly, unnatural language fragments such as syllable-only stimuli or manually synchronized speech-gesture stimuli with our without subjectively matching utterances were used. For methodological reasons, these stimuli types were fitting for their respective contexts, namely to research fused percepts or ERP responses. In order to find out how listeners in natural conversational settings perceive audiovisual speech-gesture (a)synchrony, different stimuli are required – naturally occurring and semantically whole utterances. Secondly, the direction and range of asynchronies between speech and gestures has been very limited. While gestures have a tendency to begin slightly before their co-expressive speech (e.g. Morrel-Samuels & Krauss, 1992; cf. de Ruiter, 2003), the gesture initiating after the speech is also possible, not only with deictic gestures, and especially when

videos are played or streamed. A selection of stimuli with bimodal advance and delay will be essential to studying the perception of desynchronized speech and co-produced gestures as well as large enough asynchronies to elicit an eventual breakdown of AVI. While Massaro et al. (1996), for instance, already used rather finely grained steps of asynchrony, their testing did not go beyond ± 267 ms for speech-lip stimuli. It is speculation whether a breakdown of AVI actually occurs in listeners. Özyürek et al. (2007) as well as Habets et al. (2011) researched the AVI of speech-gesture stimuli of gestural advance only, in few steps of asynchrony within a short range. No information has yet been provided on audio advance before the gestures. Thirdly, all the mentioned research has been restricted to letting participants "select" previously defined asynchronies, limiting any findings to a subset of predefined asynchronies. More specific constraints on the window of AVI for speech-gesture utterances can only be elicited by letting participants define their own preferences for AVI. The studies presented in this paper will take into account the aforementioned methodological shortcomings by using only naturally co-occurring speech-gesture utterances, testing an extended variation of speech-gesture asynchrony including audio delay and advance, and by eliciting the preferences of the participants through a rating task as well as through an active resynchronization task.

1.4. Methodological Overview

More steps of asynchrony are required to find the optimal and tolerable AVI windows of speech and gesture. The asynchronies should include delay and advance of speech to explore more possibilities. It is paramount to find out whether listeners are at all sensitive to the timing when they perceive multimodal speech because otherwise experiments on the perceived and accepted synchronies by the listeners might not produce reliable results. More information is needed on the listeners' sensitivity when it comes to the synchrony of speech and gesture in natural communication. This necessitates a methodology using natural, spontaneous language and combining identification or judgment tasks with the participants' ability to reproduce their individual preferences of simultaneity.

Two consecutive sets of studies examined naturally co-produced speech and gesture fragments. The first set of studies is an online Perceptual Judgment Task in which stimuli encompassing seven steps of asynchrony, including audio advance and delay to equal parts up

to ± 600 ms, are rated for their acceptability with varying degrees of head obscurity. It includes speech-gesture stimuli as well as physical event stimuli. The Perceptual Judgment Task will probe the windows of AVI found in previous research and inform us about the range of asynchronies for the stimuli in our second set of studies.

In the Preference Task, participants have to actively re-synchronize the audio and video channels of selected speech-gesture stimuli as well as of physical event stimuli. They use a slider to adjust the synchrony to what they feel is correct, providing us with subjective preferences of AVI. This combination of methodologies will elicit both *acceptable* as well as *preferred* windows for speech-gesture AVI on a continuous scale instead of just ratings of preselected possibilities.

2. The Perceptual Judgment Task

2.1. Study 1

Participants

141 German native speakers with mixed backgrounds completed Study 1 (mean age = 24.32, range = 16-67 years, 41 males). They rated the perceived naturalness of 2523 stimuli.

Materials

The corpus of naturalistic cartoon narrations from which we created the stimuli was already used by Kirchhof (2011), who showed that any speech-gesture co-occurrence will most likely be integrated (correctly) by the listener when a semantically complete and acceptable utterance is presented. We created 28 clips each of which contains a full utterance with a fairly prominent, naturally co-occurring gesture (7 deictic, 20 iconic, 1 emblematic; see Terminology). In Figure 1: Screenshot of the stimulus knock_1_0 containing an iconic gesture., the coding used is depicted for one of the clips: A left-handed female speaker mimics the cat Sylvester knocking on a door while saying (roughly) “where he then is dressed as a room boy an’ knocks”:

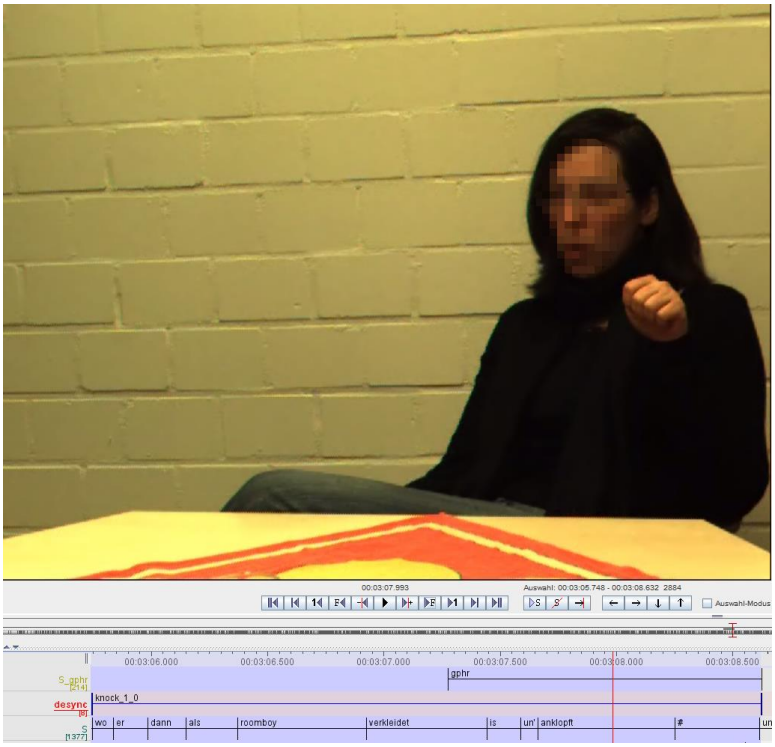

Figure 1: Screenshot of the stimulus knock_1_0 containing an iconic gesture.

The levels of audiovisual asynchrony in previous studies were restricted to small ranges and steps with a focus on video advances: Massaro et al. (1996) included audio advances and delays of 0 ms, 67 ms, 167 ms, 267 ms, and 500ms in their speech-lip stimuli, using small steps and a range of 1 s; Campbell and Dodd (1980) tested a large range of asynchronies, i.e. 3.2 s, in large steps of 400 ms, 800 ms, and 1600 ms. For speech-gesture stimuli, Habets et al. (2011) restricted their ERP testing to gestural advances of 160 ms and 360 ms, while Özyürek et al., 2007 used only the manually synchronized combination of lexical target and gesture stroke. In order to further approximate the optimal as well as the acceptable range for speech-gesture AVI, for the Peceptual Judgment Task, a) the channels were desynchronized in both directions and b) offsets in steps of 200 ms[2] up to ± 600 ms were selected to include and go beyond the previously tested time frames (see Figure 2). Whenever offsets are mentioned with regard to the studies conducted, negative offsets will indicate the speech is in delay, after the gesture (e.g., -400 ms = GS by 400 ms), while positive offsets will have the speech in advance, before the gesture (e.g., +400 ms = SG by 400 ms)

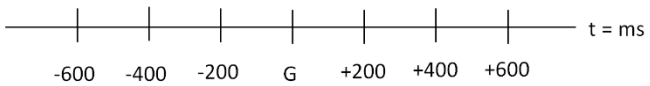

Figure 2: Scale of speech-gesture offsets.

Each of the 28 original clips was trimmed in Adobe Premiere Pro to start and end with the full gesture phrase and to include the full utterance. After the audio track was shifted in steps of 200 ms, the resulting gaps of overlap between the tracks were filled with silences and fitting still frames of the same video; both channels were contained in one and the same file. The 192 stimulus clips were put on a local server in video formats designed for different web browsers (.ogg, .mp4, .avi).

Following the Nyquist sampling theorem, to avoid aliasing, i.e. artifacts in the signal during playback, "[t]he sampling frequency should be at least twice the highest frequency contained in the signal. (…) Or in mathematical terms: $f_s \geq 2f_c$" (Olshausen, 2000, p. 1). The video containers used for our research have a frame rate of 25 fps (25 Hz), i.e. one frame every

[2] A shift by 100ms was not possible because the videos were filmed at 25 fps and Adobe Premiere Pro (CS5) does not allow for half-frame cuts. Conducting the study with asynchrony-steps per frame would have been too extensive for the participants.

40 ms. It can be assumed that any lagging of the video will not be noticeably different from a frame interval of 40 ms. The steps of asynchrony between the stimuli used in the Perceptual Judgment Task are of 200ms, which makes the intervals to be 2*5 Hz and hence well within the restrictions of the sampling theorem. The audio track has a sampling rate of 44.1 kHz; with 20 kHz being the maximal audible frequency for people with 100% hearing capability, the Nyquist sampling theorem applies with 44.1 kHz > 2*20kHz for the audio track of the stimuli used.

Procedure

A web link to Study 1 was spread via mailing lists and social media platforms (university students, Facebook, etc.). After a biographical questionnaire, participants were informed the study would take about 15 minutes and were also strongly advised to use headphones. They were told to rate the naturalness of 28 excerpts of retellings from the Canary Row cartoon in which the video or audio had sometimes been manipulated. The participants were instructed to watch the clips as often as they liked before rating them as ‘fully natural’ (‘völlig natürlich’), ‘somewhat natural’ (‘irgendwie natürlich’), ‘somewhat unnatural’ (‘irgendwie unnatürlich’), ‘fully unnatural’ (‘völlig unnatürlich’), or ‘other’ (‘sonstiges’), the latter with an option to elaborate. In a trial run with three stimuli, the participants were presented with three versions of the same original clip with an iconic gesture: One in which the audio is 1 s before the video, one with the channels in their originally recorded synchrony, and a third in which the video is 1 s before the audio. For each participant, 26 clips were selected by a script in such a way that every original stimulus would occur only once and no level of asynchrony be presented twice in a row. As in the trial run, the participants could only continue to the next clip when they had selected a rating on the scale. The judgments were recorded in an SQL database, including detailed coding of the clip variants as well as participant IDs with profiles and dropout logs. Throughout the study, a progress bar with the remaining percentage of the study was displayed.

Digression on online surveys

The usage of non-supervised surveys in the humanities is occasionally regarded as problematic for reasons such as issues with the representativeness of the population, with reliability, or with the lack of track record. Taylor (2007), for instance, states on the scientificity of

typical, serious telephone media polls: “Many of the leading peer-reviewed academic journals will not accept papers based on surveys for publication unless they have a response rate of 50% or more”. This view is probably quite common or even more so when it comes to paper-based or online surveys – a tool with which the participants are left alone for the duration of the survey. While phone polls or paper surveys allow for the calculation of a response rate, this is quite difficult with surveys distributed online. There are recordings of how many clicks a website receives. Whether these hits were intentional or incidental cannot be determined. Whether participants discontinue answering a survey is often equally recorded, a system crash or discontinuation of the answering is often not. The response rate as described by Taylor (2007) does not apply for online surveys. Holding on to this or in fact any threshold would mean we could not regard most of the polls published in the media and through politics as either scientific or meaningful.

The key to a scientifically accurate survey – online or otherwise – is to design them according to the conventional quality criteria *objectivity*, *reliability*, and *validity*. An online survey needs to be constructed in such a way that it can (a) be validated with regard to construct, content, and factors, (b) be tested for their reliability of stability and inner consistency, and (c) that they are objective in their conduct, evaluation, and interpretation (Raithel 2006: 44f.). The criteria of *objectivity*, *reliability*, and *validity* of the Perceptual Judgment Task have been tested by us via a lab replication in reduced form (see Lab replication) and by, e.g., having a between-subjects design, testing for the meeting of the proper assumptions before conducting ANOVA, analyzing cross tabulations, and other standardized statistical methods. As for the representativeness of a certain population, our online studies encompass cross sections of a general population with access to the Internet rather than the general university student subset of said population and thus present a fitting profile of the general population.

Results

The gathered data were transferred from the SQL database into SPSS. Through case selection, “other”-ratings were excluded from the statistical analysis. The categorical rating variable was coded as ordinal from 0 (‘fully unnatural’) to 3 (‘fully natural’) and entered as dependent variable into a one-way univariate ANOVA with the degrees of asynchrony as factor.

This analysis revealed a significant main effect of the degree of asynchrony on the degree of perceived naturalness [F(6,2516) = 33.47; MSE = 1.02; p < .01].

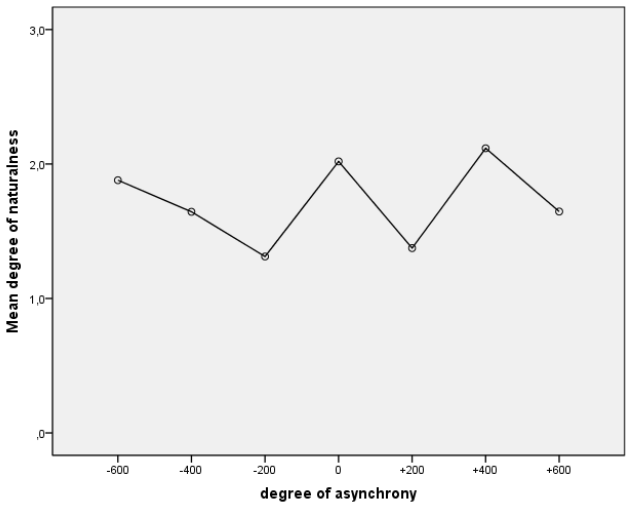

Figure 3: Mean degree of naturalness for the different degrees of asynchrony in Study 1.

The mean degree of naturalness in Study 1 was 2.287 (N = 2517, stddev = 1.049). As can be seen in Figure 3, there are peaks in the perceived naturalness at 0 ms, -600 ms and +400 ms while the stimuli desynchronized by ±200 ms are least preferred by the participants. A gradual growth in acceptance occurs between -200 ms and -600 ms of gesture before speech. The contrasts of the different levels of asynchrony with the original synchrony (0 ms) in the K Matrix show the correlations between degree of asynchrony and the perceived degree of naturalness to be significant at p < .01 for all stimuli except for those desynchronized by -600 ms (p = .064, SE = .075) and +400 ms (p = .191, SE = .075).

Discussion

The participants perceived the original condition without synchrony manipulation, an audio delay of -600 ms (GS), and an audio advance of +400 ms (SG) as most natural. The preference for the original condition, in stark contrast to the ±200 ms asynchronies, fits previous research on the McGurk effect as only asynchronies within a range of ±200 ms allow for a fused percept. The overall results of Study 1 further agree with the expected window of optimal AVI as well as with the expected breakdown of AVI, for speech-lip stimuli, beyond an asymmetric range of 500 ms (cf. Massaro et al. 1996; van Wassenhove et al. 2007). The preferred window of AVI

for speech-gesture stimuli with an audio delay between -160 ms and -360 ms found by Habets et al. (2011) is not fully confirmed by our results, but the gradual growth in acceptability between the audio at -200 ms and -600 ms before the video suggests a similar tendency.

The results confirm our methodology being appropriate to research audiovisual asynchronies by means of this instance of the Perceptual Judgment Task. The fitting of our findings with the McGurk effect further suggests that participants mostly focused on speech-lip synchrony and speech-gesture synchrony was not the major factor in Study 1.

2.2. Study 2

That the stimuli with an audio advance of +400 ms or with an audio delay of -600 ms are ranked so highly in Study 1 might be due to cues from the lip movements in the videos. Study 2 replicates Study 1 in its methodology, but the heads in the stimuli are blurred to cancel out lip visibility.

Participants

126 German native speakers (mean age = 28.28, range = 15-67 years, 42 males) participated in Study 2. They rated a total of 1812 clips for how natural they perceived those.

Materials

The heads of the speakers were covered in the 192 stimuli from Study 1 in Adobe Premiere Pro with a blurred layer following the head movements. The graphical manipulation was justified during the instruction by referring to anonymity requirements.

Procedure

The same procedure as in Study 1 applies.

Results

The data gathered in the SQL database was again exported to SPSS and “other”-ratings were removed after being checked and documented. The univariate ANOVA shows a significant contrast between the visibility conditions of Studies 1 and 2 [$F(1,4321) = 55.049$; $MSE = .97$; $p < .001$]. The ratings (3 = ‘fully natural’; 2 = ‘somewhat natural’; 1 = ‘somewhat unnatural’; 0 ‘fully unnatural’) were entered into a univariate ANOVA with the degrees of

asynchrony as factor. This analysis revealed no significant main effect of the degree of asynchrony on the degree of perceived naturalness [F(6,1805) = 1.46; MSE = .89; p = .190].

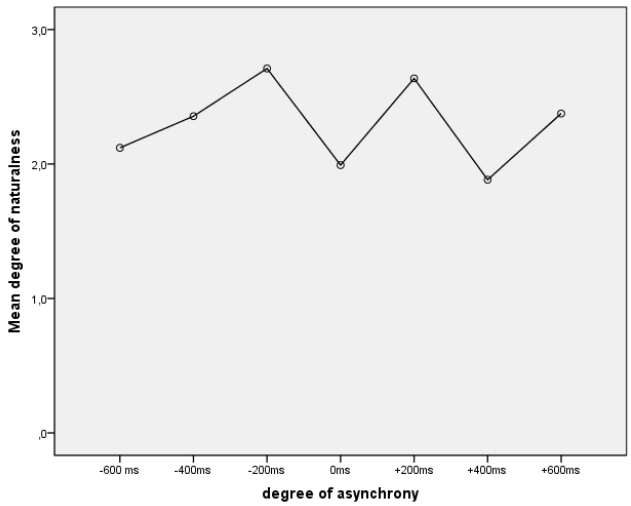

Figure 4: Mean degree of naturalness for the different degrees of asynchrony in Study 2.

The mean degree of naturalness in Study 2 was 1.937 (N = 1812, stddev = .9443). The participants rated the stimuli with asynchronies of ±200 ms as most natural (see Figure 4) and rated the original condition stimuli, which were not desynchronized, nearly as low as those with an audio advance of +400 ms. A stark contrast exists between the originally synchronous stimuli and those with an audio delay of -200 ms ($p < .05$), while the other degrees of asynchrony share a high similarity in their ratings for naturalness. Those stimuli with an audio advance of +400 ms are rated the most similar to the original condition ($p = .964$).

Discussion

There was a significant contrast for the factor 'visibility' between Studies 1 and 2 ($p < .001$). This confirms that lip visibility was an influential factor in Study 1, which leads to the similarity of the preferred windows of AVI found in previous research. The naturalness ratings in Study 2 do not replicate those from Study 1. There is no significant variation between the different degrees of asynchrony except for the participants' preference of stimuli with an audio delay of -200 ms ($p < 0.05$). This fits well with the overall tendency of previous research that audio delay is generally preferred by listeners to audio advance (cf., e.g., van Wassenhove et al. 2007; Massaro et al. 1996). An overall precedence of gesture over speech has been equally

observed in production (e.g. Thies, 2003; Morell-Samuels & Krauss, 1992; Schegloff, 1984), at least for speech-gesture pairings with strong semantic boundaries. Despite the lack of significant contrast regarding the factor ‘visibility’ as well as the factor ‘degree of asynchrony’, the results of Study 2 will be regarded as indicative of a tendency of the participants to prefer an advance of gestures before speech.

2.3. Study 3

The head motion being still noticeable in the stimuli of Study 2 might have influenced the participants in rating the different asynchronies. The blurry coverage of the speakers’ heads might have caused them to rate most stimuli as ‘somewhat natural’ because of the rather frequent usage of this type of anonymization in TV shows and newspapers. To avoid any and all visible prosodic indicators, no head movements at all are visible to the participants in Study 3, with gesture and speech remaining as the only ratable factors.

Participants

325 native German speakers (mean age = 24.31, range = 17-67 years, 85 males) rated the naturalness of 5165 stimuli in Study 3.

Materials

In Adobe Premiere, the heads from the speakers in the 192 stimuli from Study 1 were covered with a black rectangle following the head movements; motions of neither the lips nor the head are detectable by the participants. The shoulders were left uncovered to not obscure parts of the arms gesturing. The graphical manipulation was again justified by referring to anonymity requirements during the instruction.

Procedure

The same procedure as in Study 1 applies.

Results

After importing the gathered data into SPSS and fitting it for analysis, several univariate ANOVA were conducted. The test of between-subjects effects regarding the visibility condition reveals a significant difference between Studies 1 and 3 [$F(1,7674) = 38.390$; $MSE = .953$; $p <$

.001] and 2 and 3 [F(1,6963) = 8.886; MSE = .953; p < .005]. A univariate ANOVA shows a significant main effect of the degree of asynchrony on the degree of perceived naturalness [F(6,5158) = 6.282; MSE = .920; p < .01].

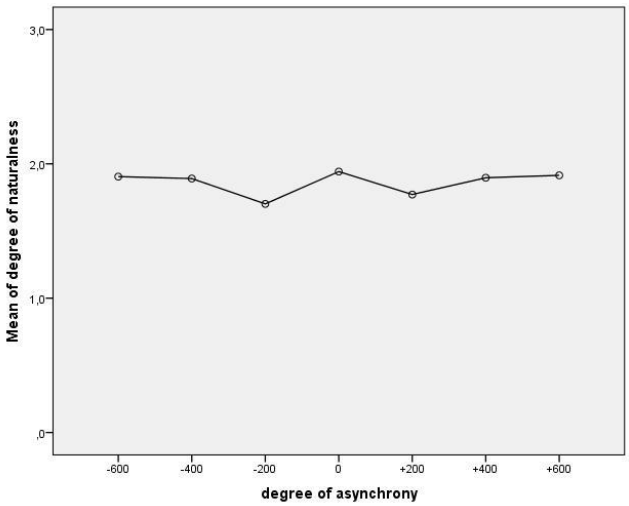

Figure 5: Mean degree of naturalness for the different degrees of asynchrony in Study 3.

The mean degree of naturalness in Study 3 is 1.860 (N = 5165, stddev = .9620), the distribution of the levels of asynchrony rated as most natural being fairly flat (kurtosis = -.847, SE = .068). The stimuli with original synchrony are clearly preferred to those with audio advances of +200 ms (p < .001) or delays of -200 ms (p < .001), as can also be seen in Figure 5. No significant contrasts were found for other levels of asynchrony with the original clips.

Discussion

Massaro et al. (1996) set the preferred window of AVI for syllables within maximal ranges of 150 to 250 ms of asynchrony and suspect a significant breakdown in the perceptual alignment at discrepancies between 250 ms and 500 ms of asymmetric asynchrony. The participants in Study 3 clearly preferred the original synchrony (0 ms) of the stimuli to the ±200 ms asynchronies (p < .001), even though about two thirds of all stimuli were rated as somewhat or fully natural. Whether this above-chance rating speaks against a breakdown of AVI or for it is debatable. This agrees with the findings by Habets et al. (2011), who hypothesize the window of AVI for single words with gestures to extend to somewhere between an auditory delay of -160 ms and -360 ms after the gestures. Our findings expand this possible window of AVI to audio advances of up to +200 ms. That all visual prosodic influence was cancelled out by the head

blockage is an argument supporting a wider window of AVI for speech with gestures than for speech alone.

2.4. Lab replication

We conducted a partial replicate of Studies 1-3 at a PC in our Nat.CoMM/HD lab for testing the studies' reliability (see Digression on online surveys).

Participants

17 participants (mean age = 25, range = 22-42 years, 6 males) rated a total of 765 stimuli, 255 in each visibility condition, with regard to how natural they perceived them.

Materials/ Procedure

After completing the same trial as in Studies 1-3, the participants were presented with three versions of one video clip in the lips-visible condition at -600 ms gesture before speech, 0 ms original synchrony, and +200 ms speech before gesture. Apart from making the lab replication more time-efficient, these degrees of asynchronies were selected because they include the suspected window of optimal AVI as well as an asynchrony at which AVI should definitely have broken down. Having watched the three stimuli several times, the participants were to choose the one most natural to them or to indicate that they were unable to decide. This procedure was then repeated for another four sets of three stimuli in the lips-visible condition (Study 1), and for another five sets of three stimuli each in the face-blurred (Study 2) and face-blocked condition (Study 3).

Results

After the data was cleaned from "undecided" ratings, a univariate ANOVA resulted in a significant main effect of visibility on the perceived naturalness of the stimuli [$F(2,208) = 3.881$; $MSE = 0.9$; $p < .05$]. Contrasts reveal the difference between the lips-visible condition and the face-blurred condition as not significant ($p = .902$), while the lips-visible condition differs significantly ($p < .05$) from the face-blocked condition, which, in turn, differs significantly ($p < .05$) from the face-blurred condition.

The participants preferred the 0 ms stimuli and the +200 ms audio advance stimuli in the lips-visible condition while the audio delay of -600 ms was not selected at all (c.f. Figure 7). As can be seen from the cross tabulation (χ^2: $p < .00$) in Figure 6, the -600 ms asynchronies were clearly dispreferred across conditions (8.5%), while the original synchrony (46.9%) was nearly as preferred as the audio advance of +200ms (44.5%).

preferred degree of asynchrony * condition Crosstabulation

			condition			Total
			lips visible	blurred	blocked	
preferred degree of asynchrony	-600	% of Total	0,0%	2,4%	6,2%	8,5%
		Std. Residual	-2,6	-,2	2,9	
	0	% of Total	20,9%	16,1%	10,0%	46,9%
		Std. Residual	1,3	,6	-2,0	
	+200	% of Total	15,6%	12,3%	16,6%	44,5%
		Std. Residual	-,2	-,5	,8	
Total		% of Total	36,5%	30,8%	32,7%	100,0%

Figure 6: Cross tabulation of preferred degrees of asynchrony by visibility condition in lab replication study.

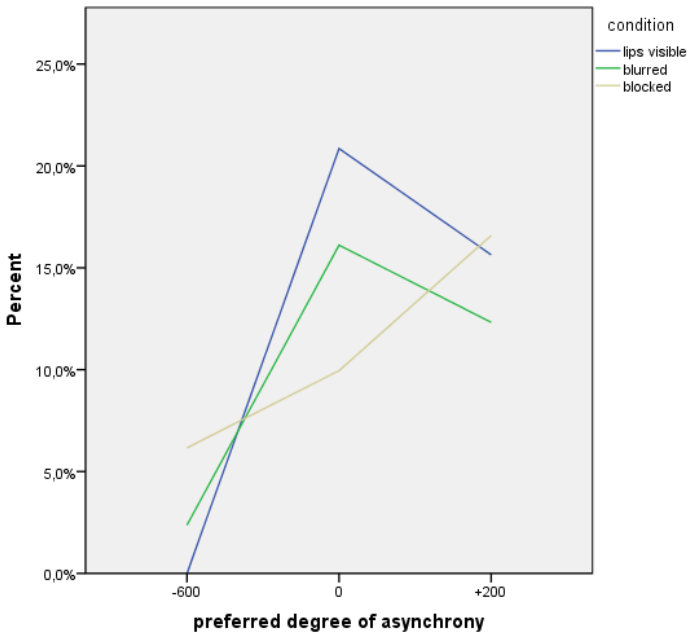

Figure 7: Preferred degrees of asynchrony in lab replication study.

Discussion

The lab replication study supports the findings from Studies 1-3 and thus confirms our methodology. As in Study 1, the original 0 ms synchrony is slightly preferred to the audio advance of +200ms, while no audio delay is rated as natural by any participant. As the head visibility decreases, the preference distribution among the degrees of asynchrony increases. The perceived naturalness of the original synchrony is less accepted in the face-blocked condition (see Study 3), the audio delay of -600 ms has at least minimal acceptability. Finally, as can be clearly seen in Figure 4, the medial naturalness ratings even out between the original synchrony and the audio advance of +200 ms, which is comparable to the flat distribution among the naturalness ratings in Study 3 (cf. Figure 5).

2.5. Study 4 (physical)

The lab replication study has verified the reliability of the Perceptual Judgment Task. The factor of lip synchrony and prosodic head movements has been eliminated during the course from Study 1 to Study 3 and wider windows of AVI for speech-gesture stimuli turn have become more likely than previously assumed. Study 4 will provide acceptable audiovisual asynchronies from stimuli we know to be causally and temporally fixed as a baseline to be compared to the speech-gesture ratings.

Participants

142 participants (mean age = 27.86, range = 18-62 years, 40 males) rated 2249 physical cause-and-effect stimuli in Study 4 of the Perceptual Judgment Task.

Materials

In Adobe Premiere, ten short videos of physical cause-and-effect stimuli were desynchronized into the previously used seven degrees of asynchrony. The original clips contained exactly one instance of each of the following: a hammer hitting a nail, snapping a book shut, a clap of the hands, clinking a class with a fork, a tap on a keyboard, knocking on a table, the plop while opening a bottle of champagne, fingers snapping, hitting a bass drum, and

popping a balloon with a needle. While the hammer hitting the nail functioned as a trial stimulus, the other nine sources of noise were used in the actual study.

Procedure

The same procedure as in Study 1 applies.

Results

After importing the gathered data into SPSS and fitting it for analysis, a univariate ANOVA revealed a significant main effect of the degree of asynchrony on the degree of perceived naturalness [$F(6, 2248) = 71.966$; $MSE = 1.046$; $p < .01$].

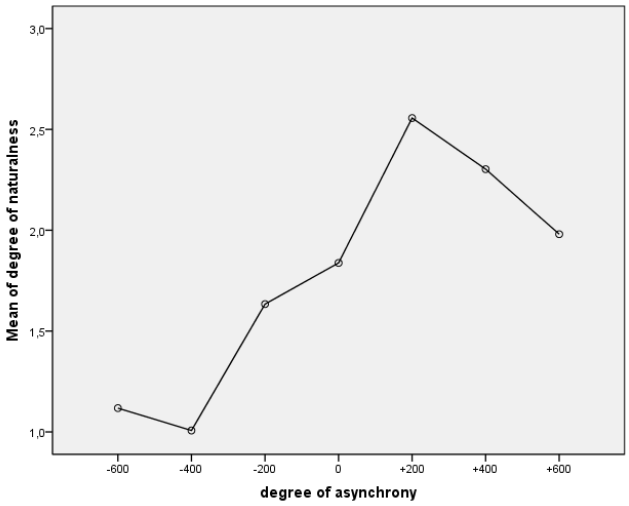

Figure 8: Mean degree of naturalness for the different degrees of asynchrony in Study 4.

The mean degree of naturalness in Study 4 was 1.724 (N = 2249, stddev = 1.1155), and already from the means graph we can see a clear difference between the perceived naturalness of the speech-gesture stimuli and the physical event stimuli. While in all visibility conditions, the graph peaked at different levels of asynchrony between speech and gestures, in Study 4 the participants distinctly preferred the physical event stimuli in which the audio precedes the video by +200 ms (Figure 8). This is confirmed by contrasts of the different levels of asynchrony in the K Matrix, which shows the audio advance of +200 ms to be significantly different at $p < .01$ for all levels of asynchrony but for +600 ms, which is still significantly different at $p < .05$.

The distribution of the preferred levels of asynchrony in Study 4 is barely skewed and rather platykurtic (skewness = -.271, kurtosis = -1.299, SE = .103) around the mean of 1.72 on a scale of 0 (fully unnatural) to 3 (fully natural).

Discussion

The window of optimal AVI for physical bimodal stimuli as determined by van Wassenhove et al. (2007) ranges from -200 ms to +200 ms around the original audiovisual synchrony, which is a slightly smaller range than the 533 ms Massaro et al. (1996) suggested for the possible window of AVI for speech-lip signals before integration breakdown. The participants in Study 4 displayed a clear preference for stimuli in which the audio precedes the video by +200 ms, which goes in line with previous research on the AVI of bimodal media in psychophysics (also see Section 1.2). As with the speech-gesture stimuli in Studies 1-3, the audiovisual synchrony of cause-and-effect stimuli has a significant effect on the naturalness as perceived by the participants. The well-formed distribution of the ratings provides additional support of our methodology. The results of Study 4 hence provide an excellent baseline to which we can compare the preferred asynchronies of the speech-gesture stimuli.

2.6. Discussion Perceptual Judgment Task

The aim of Studies 1 through 4 in the Perceptual Judgment Task was to find an optimal and tolerable window of AVI for co-occurring speech and gesture utterances on the basis of previous research on the perception of audiovisual signals in general and on speech-lip synchrony specifically. The findings by Massaro et al. (1996) and van Wassenhove et al. (2007), among others, lead to suspect that the participants in the Perceptual Judgment Task would prefer audiovisual asynchronies between ±200 ms, while Habets et al. (2011) and Özyürek et al. (2007) found preferred windows of AVI for speech-gesture combinations between -160 ms and -360 ms of speech after the gesture specifically. To narrow down the optimal and tolerable window of AVI for co-occurring speech and gesture utterances, we used more extensive degrees of asynchrony and expanded our research to include delay and advance of both audio and video signals in equal shares. Another novelty in our methodology was the usage of naturally co-occurring speech-gesture utterances (Studies 1-3) and physical cause-

and effect stimuli (Study 4) rather than artificially combined audiovisual signals in order to make more reliable predictions on natural perception.

The results of Study 4 can be seen as confirmation of the judgment ability of the participants in Studies 1 through 3. They were able to discern between asynchronies differing by steps of ±200ms and to rate these asynchronies with regard to how natural they perceived them. Due to the full lip visibility in Study 1, this data could be compared with the findings from previous research on the AVI of speech-lip stimuli (cf. Results). Study 2 presented the participants with reduced versions of the stimuli from Study 1 – while the lips were hidden by blurring out speakers' faces, head motion was still visible. The results of Study 2 reflect those of Study 1 in terms of the ratings of naturalness for the different degrees of asynchrony; the effect of the degree of asynchrony on the degree of perceived naturalness was not significant in Studies 1 and 2 (cf. Results). In Study 3, the stimuli with original synchrony were clearly preferred to those with audio advances of +200 ms ($p < .001$) or delays of -200 ms ($p < .001$), whereas no significant contrasts could be found for the other levels of asynchrony. That participants were not able to clearly rate the perceived naturalness of speech-gesture asynchronies beyond ±200 ms supports the findings from previous research on the optimal window of AVI for such stimuli. A breakdown in the perceptual alignment at discrepancies between ±250 ms and ±500 ms in either direction, as suggested by Massaro et al. (1996) for multimodal syllable stimuli, seems likely for speech-gesture stimuli as well.

3. The Preference Task

Since all stimuli with obscured heads received naturalness-ratings of more than 60% in studies 2 & 3 of the Perceptual Judgment Task, no specific temporal window of AVI for speech and gesture can be estimated on the basis of these results. The window might go beyond the presented levels of asynchrony or it might lie somewhere in between. The Preference Task, using a different methodology, is designed to further approximate the possible range of AVI for speech-gesture utterances. It is aimed at finding out whether listeners can *reproduce* the production synchrony; whether they can specify what timing of speech and gesture they prefer without being given options to choose from. Study 5 examines the stimuli from the Perceptual Judgment Task to investigate possible differences in the AVI of speech and gestures in general. Study 6 then focuses on the variation of AVI between different gesture types.

3.1. Study 5

Participants

20 native speakers of German took part in Study 5 (mean age = 25.80, range = 21-40 years, 6 males). The participants were all university students and the two best performers, i.e. those who got closest to the original synchronizations, were promised a 25€ voucher for a popular online retailer. This incentive was intended to make the participants more motivated in the tedious task of re-synchronizing the stimuli.

Materials

The test of between-subjects effects for head visibility in the Perceptual Judgment Task revealed a significant difference between Studies 1 and 3 [$F(1,7674) = 38.390$; $MSE = .953$; $p < .001$] and Studies 2 and 3 [$F(1,6963) = 8.886$; $MSE = .953$; $p < .005$]. To ensure no prosodic visual cues distract participants from the gestural stimulus, the variants with the blocked-out heads from Study 3 were used for the Preference Task. 15 clips with prominent iconic gestures were selected and manipulated with five different initial asynchronies, dependent on the frames in Adobe Premiere (277 ms, 451 ms, 607 ms, 754 ms, 1034 ms). Silences stemming from the original recordings were put in front of the audio to create fragments of equal length as the video tracks. The resulting clips were expanded before and after the fragments with silences

and still frames to create space for “sliding” the channels back and forth during the resynchronization. The interface for re-aligning the audio (.mp3) with the video (.mov) was the annotation program ELAN (Wittenburg, Brugman, Russel, Klassmann, & Sloetjes, 2006) 3.9.1 in its media-synchronization-mode.

In order to verify the methodology of the slider study and at the same time test the participants’ AVI-abilities, we also desynchronized two physical events from Study 4 in the same manner as the gesture clips. The stimuli of a hammer hitting a nail and of fingers snapping were each manipulated to have the video precede the audio by 902 m. This strong asynchrony was selected to avoid participants accepting the desynchronized video as the original. The physical event stimuli were used as a trial and also functioned as baseline for the speech-gesture stimuli.

Procedure

Study 5 was conducted in a quiet room on a notebook (1366x768p; 15.6") with the sound coming from closed headphones (Sennheiser HD 201). The 15 stimuli were given in reversed order to half of the participants to control for sequential effects and the video size, screen contrast, and brightness were kept constant. The instructor explained the ELAN synchronization interface[3] to the participants and showed them with the help of an example stimulus how resynchronize the channels by sliding the audio offset. In the interface, the audio and video channels are accessed through two media players. With the extended video track being fixed, the participants were able to “slide” the audio file into place onto the video track (Figure 9). The participants’ task was to resynchronize the clips until they felt they were synchronized correctly. They were encouraged to prioritize precision in their resynchronization.

[3] http://www.mpi.nl/corpus/manuals/manual-elan.pdf

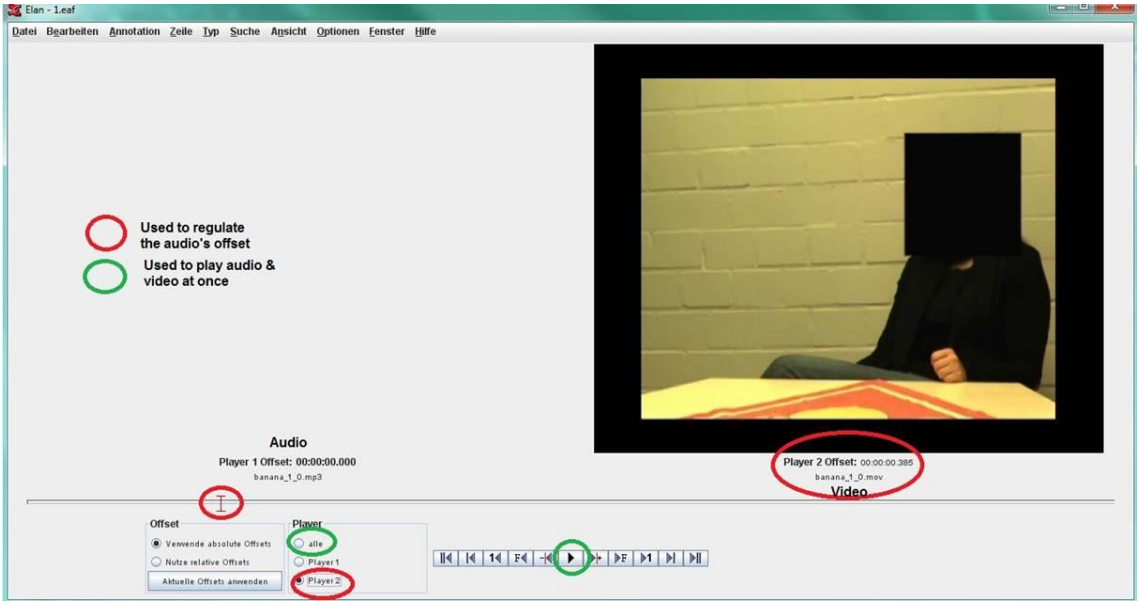

Figure 9: ELAN in synchronization mode as used by the participants of the Perceptual Judgment Task to resynchronize the audio and video of the stimulus shown in Figure 1.

Results

The asynchronies as set by the participants were entered into a spreadsheet program and the divergences from the original clip synchrony were calculated. This way, the actual preferred offsets from the original synchrony were determined. After transferring the calculated data into SPSS, descriptive statistics were elicited.

The asynchronies set for the physical cause-and-effect stimuli had a range of 1420 from -978 ms of audio delay to +442 ms audio advance (excluding outliers 1 through 4: 440ms; -154 ms to +286 ms) with a mean of 13.18 ms (N = 40, stddev = 245.495). The asynchronies set by the participants for the speech-gesture stimuli had a range of 2662 ms from -1908 ms audio delay to +754 ms audio advance (2014 ms; -1361 ms to +653 ms excluding outliers 1-4) with a mean of -72.59 ms (N = 300, stddev = 421.327). This difference in range and variation is clearly displayed in Figure 10 and Figure 11.

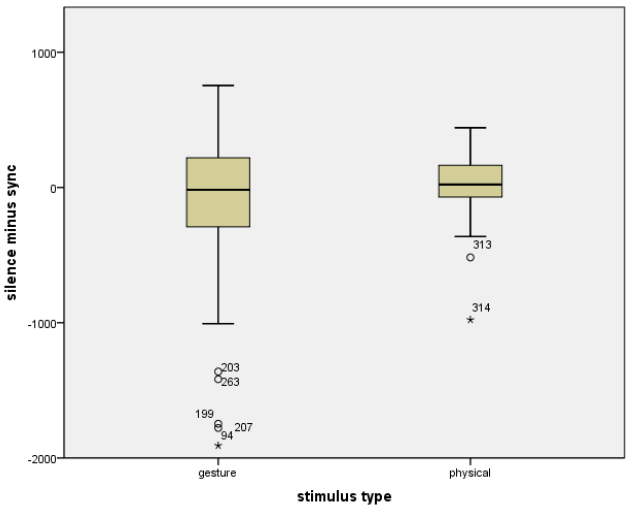

Figure 10: Range of asynchronies set for different stimulus types in Study 5.

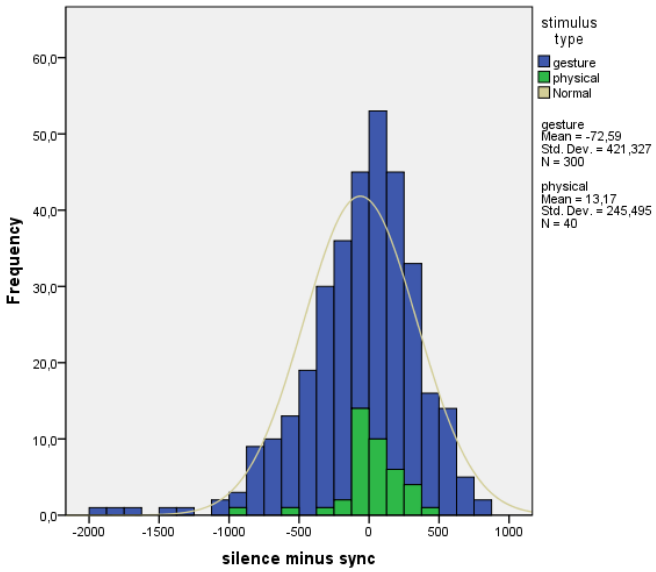

Figure 11: Histogram of range of asynchronies set for different stimulus types in Study 5.

The data were entered into a univariate ANOVA, which revealed no significant main effect of the stimulus type on the preferred asynchrony of the stimuli [F(1,338) = 1.5836; MSE = 163987.753; p = .209].

Discussion

The ranges of asynchronies set by the participants in Study 5 varied greatly for the physical cause-and-effect stimuli (440 ms) and the speech-gesture stimuli (2014 ms). Not only is the range for the speech-gesture stimuli wider, but also is the variation within this range. The

participants were able to approximate the preferred asynchronies for physical cause-and-effect stimuli from the Perceptual Judgment Task (> -200 ms and < +400 ms). The optimal window for AVI in Study 5 for the physical stimuli, excluding outliers, lies between -154 ms of audio delay and declines over +286 ms of audio advance towards a cutoff at +442 ms.

As in the online Preference Task, the participants in Study 5 displayed a wider range of acceptance for the speech-gesture than for the physical stimuli. The window of AVI set by the participants is distributed in a near-normal curve around the +200 ms mark and spreads out rather evenly between an audio delay of -1361 ms and an audio advance +653 ms, excluding outliers (See Figure 11). While Habets et al. (2011) and Özyürek et al. (2007) found preferred windows of AVI for speech-gesture combinations between -160 ms and -360 ms of speech delay, the results of Study 5 clearly broaden these preferred windows. And, even though the data does not support the breakdown of AVI for discrepancies between ±250 ms and ±500 ms as suspected by Massaro et al. (1996) for speech-lip syllable stimuli, their hypothesis of a window for optimal AVI between ±250 ms still holds.

3.2. Study 6

Study 5 showed the slider methodology to be appropriate for eliciting the preferred audiovisual synchronization of our participants for speech gesture stimuli as well as for physical cause-and-effect stimuli. While it made use of mostly iconic and iconic-metaphoric gestures, the speech-gesture continuum McNeill (2005, p. 7) described based on Kendon (1988) suggests that a variation in temporal synchrony preference might apply for different gesture types, namely for emblems versus other “gesticulations”. Study 6 examines this possible variation using the methodology of Study 5 for selected new stimuli.

Participants

23 German native speakers (mean age = 27.91, range = 20-45 years, 10 males) completed the Preference Task in Study 6. They were again gathered from the university student population and an incentive was provided to enhance their motivation for precision.

Materials

6 different physical cause-and-effect events not previously used as well as 13 novel speech-gesture stimuli, i.e. 4 deictic, 3 emblematic, and 6 iconic gestures (see McNeill 2005:38f.) were created using the same methodology as in Study 5.

Procedure

The participants were presented with the same experimental setup as in Study 5, each being instructed to resynchronize the 6 physical-event clips for means of a trial, and the 13 speech-gesture clips as the actual study.

Results

The resynchronization of the physical cause-and-effect stimuli resulted in a range of 1542 ms between an audio delay of -966 ms and an audio advance of +576 ms (excluding outliers 1 through 4: 881 ms; -597 ms to +284 ms) with a mean of -121.61 ms (N = 144, stddev = 228.504). The range of preferred synchronization varies for speech-gesture stimuli (see also Figure 12 and Figure 13). The overall range in which the participants resynchronized the speech-gesture stimuli is along 3955 ms, between an audio delay of -2921 ms and an audio advance of +1034 ms (excluding outliers 1 through 4: 1534 ms; -927 ms to +607 ms) with a mean of 39.14 ms (N = 312, stddev = 360.720).

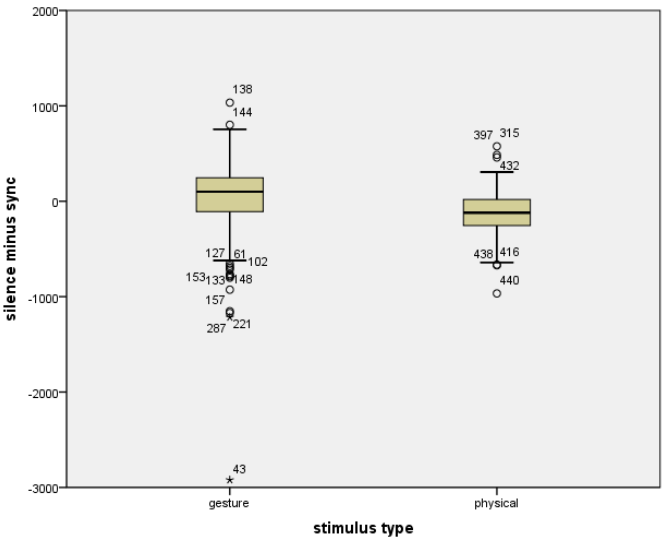

Figure 12: Range of asynchronies set for different stimulus types in Study 6.

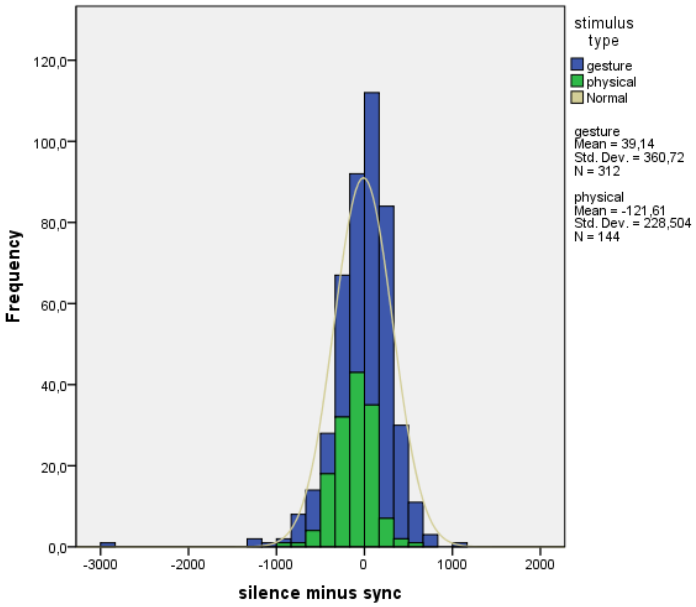

Figure 13: Histogram of range of asynchronies set for different stimulus types in Study 5.

The three gesture types tested vary from this overall range as follows (see also Figure 14 and Figure 15): The iconic gestures were aligned with their coproduced speech by the participants along a range of 3955 ms between -2921 ms of audio delay and an audio advance of +1034 ms (excluding outliers 1 through 4: 1249 ms; -655 ms to +594 ms) with a mean of 30.17 ms (N = 144, stddev = 395.233).

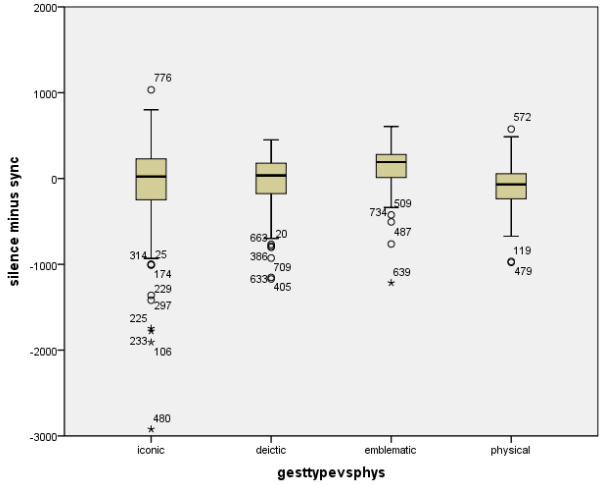

Figure 14: Range of asynchronies set for different gesture type and physical events in Study 6.

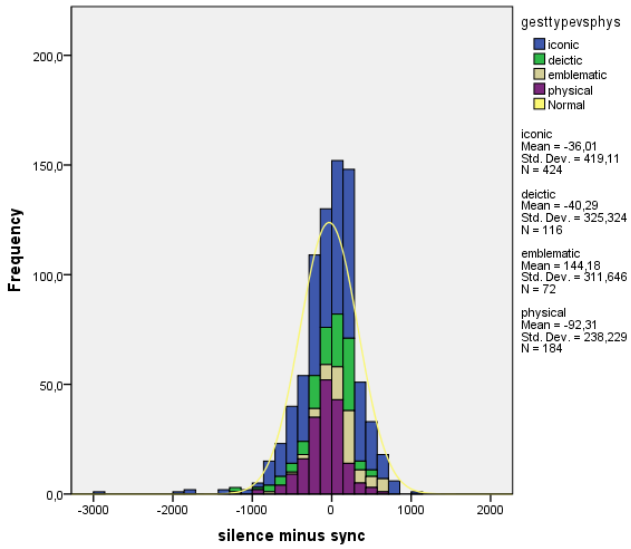

Figure 15: Histogram of range of asynchronies set for different gesture type and physical events in Study 6.

The deictic speech-gesture stimuli were resynchronized along a range of 1622 ms between an audio delay of -1171 ms and an audio advance of +451 ms (excluding outliers 1 through 4: 1141 ms; -787 ms to +354 ms) with a mean of -26.17 ms (N = 96, stddev = 324.732). The emblematic gestures were realigned with their coproduced speech by the participants along a range of 1823 ms between an audio delay of -1216 ms and an audio advance of +607 ms (excluding outliers 1 through 4: 942 ms; -337 ms to +605 ms) with a mean of 144.18 ms (N = 72, stddev = 311.646).

All data from Study 6 was recoded to run an ANOVA with the different gesture types versus the physical stimuli as factor. This analysis revealed a significant main effect of this variable on the synchrony set by the participants [$F(3,455) = 12.131$; $MSE = 103358.691$; $p < .01$]. Contrasting the different gesture types with the physical events, the iconic and emblematic gestures were highly different ($p < .01$). The deictic gestures elicited slightly less yet significant temporal differences in the participant synchrony preferences ($p < .05$).

Discussion

The preferred overall speech-gesture synchrony in Study 6 had a range of 1534 ms (-927 ms to +607 ms) while the preferred physical cause-and-effect synchrony ranged over 881 ms (-597 ms to +284 ms). Apart from the general difference in preferred synchrony, the tendency of the participants to select an audio advance in the stimuli is prominent. This goes in

line with light traveling faster than sound (see, e.g. Einstein, 1905/2005)[4] and hence asynchronous production facilitating a synchronous perception.

The striking finding of Study 6 is the variation of preferred audiovisual synchrony for different types of gestures with their coproduced speech. Expanding on McNeill (2005, p. 7), speech and gesture should be more closely semantically linked for iconics than for deictics, which is also reflected in the temporal synchrony in our data (iconic gestures 1249 ms; -655 ms GS to +594 ms SG vs. deictic gestures 1141 ms; -787 ms GS to +354 ms SG). In the same continuum of semantic synchrony, emblems are described as least semantically linked to speech since they are comprehensible without speech. In Study 6, we examined emblematic gestures with naturally co-occurring redundant speech, which resulted in the closest preferred temporal synchronies of all gestures (emblematic gestures 942 ms; -337 ms GS to +605 ms SG). This in turn might be due to the tight semantic affiliation, which is less present in deictic speech-gesture combinations and mostly associative in iconic speech-accompanying gestures. The smaller window of AVI for emblematic and deictic gestures with co-produced speech is closer to the preferred window of AVI for physical cause-and-effect stimuli (881 ms; -597 ms GS to +284 ms SG). While speech is not caused by gestures in the same way it is caused by air flow through the speech apparatus, there are certain multimodal proximity pairs expected by the listener to occur together, such as a deictic verbal expression like “over there” with a gestural one like pointing over there alongside it. An even stronger expectation of semantic alignment, with our without temporal synchronization, might happen with gestural emblems – if they are accompanied by any speech at all, it should reinforce the gesture and hence be semantically redundant, such as a thumbs-up with a simultaneous “Well done!”.

3.3. Discussion Preference Task

Results Studies 5 & 6

The range of the preferred gesture-type independent speech-gesture synchrony slightly increases when the data from both studies are combined, excluding outliers, to 2172 ms (-1418 ms GS to +754 ms SG). The range of preferred physical cause-and-effect synchrony also

[4] “That’s why some people appear bright until you hear them speak” (allegedly said by Albert Einstein).

slightly increases to 995 ms (-643 ms GS to +352 ms SG). The crucial difference of more than 1 s in the preferred AV synchrony by the participants remains just as clear (see also Figure 16).

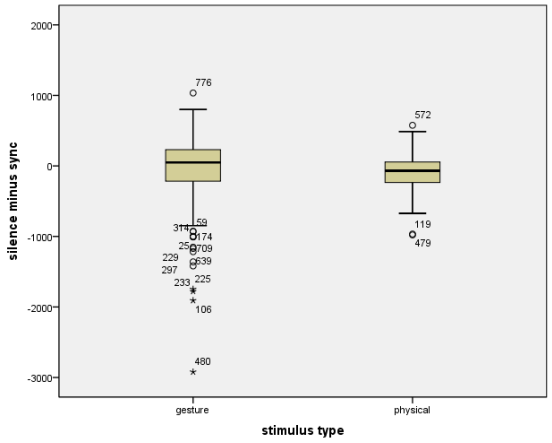

Figure 16: Range of asynchronies set for different gestures and physical events in Studies 5 & 6.

An overall main effect of stimulus type on the degree of synchrony entered by the participants was discovered [F(3,792) = 7.423; MSE = 1.03E8; p < .01]. Repeating the ANOVA for the different gesture types against the physical event stimuli returned a significant impact of the stimulus type on the preferred synchrony only for the emblematic gestures (p < .01). The iconic (p = .078) and deictic (p = .226) gestures with their co-produced speech did not contrast as strikingly with the cause-and-effect stimuli in the K Matrix. Taking the iconic gestures as the reference category, emblematic function significantly influences the preferred speech-gesture synchrony (p < .1), but deictic gestures are just about not significantly different from the iconic ones (p = .078). There is a clear variation in synchrony range between gesture types and between physical eventsFigure 17:

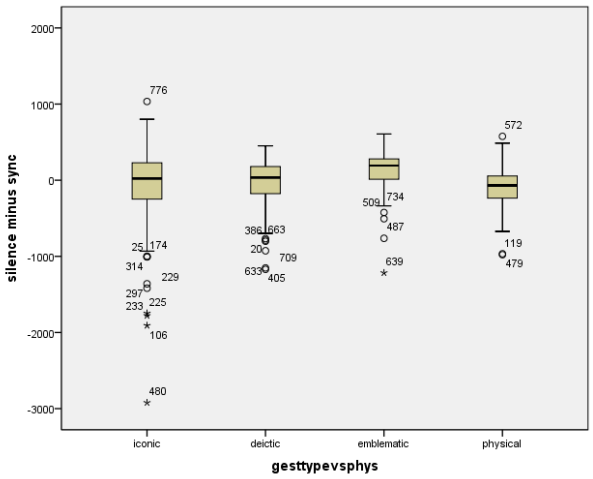

Figure 17: Range of asynchronies set for different gesture types & physical events in Studies 5 & 6.

Combining the results of Studies 5 and 6, we still have a narrow window for the preferred synchrony of physical events (-87 ms to +672 ms; MSE = 86; stddev = 214.4), and the iconic gestures are synchronized only loosely with their speech (-908 ms GS to +778 ms SG; MSE = -4; stddev = 386.4). The resynchronizations of emblematic and deictic gestures show different patterns: Both got resynchronized closer to their original timing than the iconic gestures. The deictics were readjusted more similarly to the physical events (-51 ms GS to +1171 ms SG; MSE = -5.5; stddev = 321.2), with more of a tendency toward an audio advance. The emblematic gestures were also resynchronized more closely with their non-obligatory speech (-607 ms GS to +1216 ms SG; MSE = -41; stddev = 284.4) than the iconic ones to their disambiguating speech. It appears there are some conditions for speech-gesture AVI after all.

4. General Discussion/Conclusion

The results for the physical events and speech-gesture utterances show that participants accept delays or advances in both the acoustic and the visual modality. This has been a major gap in previous research. The Preference Task supports the results of the Perceptual Judgment Task by confirming and even expanding the wide range of accepted offsets. While audiovisual stimuli such as physical events and speech-lip signals require a production-like, tight synchrony, the relevance of such a synchrony between speech and gesture is not supported by our results. Deictic and emblematic gestures do seem to entail a closer temporal synchrony to their co-occurring speech than iconic gestures. This may be due to a closer semantic relation between the modalities during the phase of synchronous production.

The audio and video in the physical events stand in a causal relationship while speech and gesture share a semantic, conceptual connection. In multimodal language production, they temporally align to a certain degree. The speech-gesture continua by McNeill (2005, p. 7ff. based on Kendon, 1988) gives a more specific explanation of the different levels of gesture-speech entrainment. McNeill classifies gestures along a continuum according to the obligatoriness of speech: for ‘gesticulations’, such as iconic and deictic gestures, speech is mandatory while for emblems it is optional; for pantomime and sign language speech need not be present. One can modify this continuum to include deictic and iconic gestures in lieu of the encompassing gesticulations:

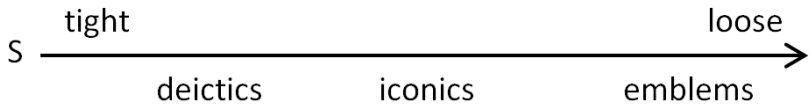

Figure 18: Continuum of semantic synchrony of speech and gesture types.

We can hypothesize that with loosening *semantic* synchrony the need for *temporal* synchrony becomes smaller because of the decreasing disambiguating function of co-occurring speech. Another factor is the theme-rheme frame discussed in Kirchhof (2011), which binds the gesture to a certain sentential and hence temporal frame of an utterance. These frames are

present in the stimuli of both the Perceptual Judgment Task and the Preference Task and the participants accepted larger temporal asynchronies than can be found in production. We hypothesized that gestures only need to synchronize loosely with their co-occurring speech. The Preference Task disproves this to a certain degree because different windows of AVI are accepted by the participants for different gesture types: emblems seem to need more synchrony with speech than deictics than iconics. This information can provide us with a sketch of a temporal continuum (Figure 19) diverging from the semantically governed one:

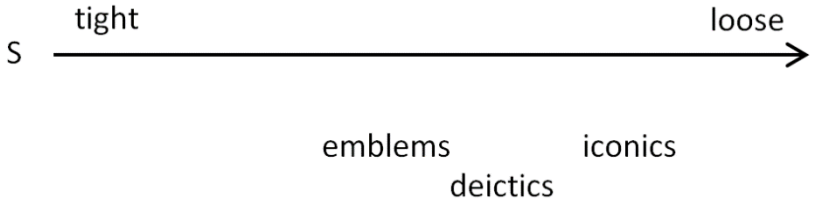

Figure 19: Continuum of temporal speech-gesture synchrony in perception.

The close temporal synchrony between speech and gesture is a well-known production phenomenon, and it seems be more important for AVI than previously thought. Since iconic gestures complement phrases and utterances, the temporal window for their AVI is only bound by the utterance duration and the timing within this boundary is very flexible. Deictic gestures correspond to deictic POS, the closest a gesture can get to lexical affiliation with speech. They are semantically and temporally bound and their phases are short, which makes the temporal window for AVI small. Emblematic gestures, then, are a special case. When they occur together with speech they are redundant to certain parts of speech (POS). In the Preference Task, participants synchronized them closely to their temporal production synchrony, which suggests a tight semantic and temporal bound between the two modalities for this gesture category. As with deictic gestures, their phases are short, but, due to their redundancy, the window for AVI is slightly larger.

As de Ruiter (2000) and Kirchhof (2011) already suggested, the relation between gestures and speech is governed by conceptual bounds. For perception, this conceptual package is transmitted by an internal (re)synchronization of the duration of the gesture phrase with the speech it is semantically associated with, by AVI. Within a theme-rheme frame,

production-like synchrony is not necessary for the listener: We suggest that gesture-speech synchrony is a predominantly production-based phenomenon. This explains why in the Perceptual Judgment Task and the Preference Task there was a wide range of accepted as well as of preferred asynchronies between the speech and co-expressed gesture: Listeners do not require speech-gesture synchrony and hence cannot reproduce it.

As McNeill (2012) speculated on the conceptual transmission of a speech-gesture utterance, “the time limit on growth point asynchrony is probably around 1~2 secs., this being the range of immediate attentional focus” (33). The GP is temporally very flexible in perception, with the possibility of either modality preceding the other by up to 1418 ms, depending on the gesture type. We can observe a semiotic connection between the two modalities by analyzing co-produced speech and gestures. What we cannot do is easily desynchronize or semantically mismatch speech and gesture during production (cf. Holler et al., 2009). Our results strongly suggest s speech-gesture synchrony is rather a consequence of the production system but, as far as actively set preferences are concerned, seems not to be crucial for comprehension. This finding should allow for a higher tolerance of timing in modeling gestures in virtual agents and robots and could inform and inspire future research into the perception of naturally co-occurring speech and gestures.

5. References

Alibali, Martha W., Dana C. Heath, & Heather J. Myers (2001). Effects of visibility between speaker and listener on gesture production: Some gestures are meant to be seen *Journal of Memory and Language, 44* (2), 169-188.

Bergmann, Kirsten & Stefan Kopp (2009). Increasing expressiveness for virtual agents - Autonomous generation of speech and gesture for spatial description tasks. In Keith S. Decker, Jaime S. Sichman, Carles Sierra, & Cristiano Castelfranchi (Eds.), *Proceedings of the 8th International Conference on Autonomous Agents and Multiagent Systems* (pp. 361-368). Ann Arbor, MI: IFAAMAS.

Bushara, Khalafalla, Jordan Grafman, & Mark Hallett (2001). Neural correlates of auditory-visual stimulus onset asynchrony detection. *Journal of Neuroscience*, *21*, 300-304.

Callan, Daniel E., Jeffery A. Jones, Kevin Munhall, Christian Kroos, Akiko M. Callan, & Eric Vatikiotis-Bateson (2004). Multisensory-integration sites identified by perception of spatial wavelet filtered visual speech gesture information. *Journal of Cognitive Neuroscience*, *16*, 805-816.

Cassell, Justine, David McNeill, & Karl-Erik McCullough (1999). Speech-gesture mismatches: Evidence for one underlying representation of linguistic and non-linguistic information. *Pragmatics and Cognition*, *7* (1), 1-33.

De Ruiter, Jan P. & David P. Wilkins (1998). The synchronization of gesture and speech in Dutch and Arrernte (an Australian Aboriginal language): A cross-cultural comparison. In Serge Santi (Ed.), *Oralité et gestualité* (pp. 603-607). Paris: L'Harmattan.

De Ruiter, Jan P. (2000). The production of gesture and speech. In David McNeill (Ed.), *Language and Gesture* (pp. 284-311). Cambridge, UK: CUP.

Einstein, Albert (1905/2005). Zur Elektrodynamik bewegter Körper. *Annalen der Physik und Chemie*, *17*, 891-921.

Feyereisen, Pierre (2007). How do gesture and speech production synchronise? *Current Psychology Letters: Behaviour, Brain and Cognition*, *23* (2), 2-12.

Fujisaki, Waka & Shin'ya Nishida (2005). Temporal frequency characteristics of synchrony-asynchrony discrimination of audio-visual signals. *Experimental Brain Research*, *166*

(3), 455-464.

Gullberg, Marianne & Kenneth Holmqvist (2006). What speakers do and what listeners look at. Visual attention to gestures in human interaction live and on video. *Pragmatics and Cognition*, *14*, 53-82.

Gullberg, Marianne & Sotaro Kita (2009). Attention to speech-accompanying gestures: Eye movements and information uptake. *Journal of Nonverbal Behavior*, *33* (4), 251-277.

Gut, Ulrike. (2009). *Non-native speech: A corpus-based analysis of phonological and phonetic properties of L2 English and German*. Frankfurt: Peter Lang.

Habets, Boukje, Sotaro Kita, Zeshu Shao, Asli Özyürek, & Peter Hagoort, (2011). The role of synchrony and ambiguity in speech-gesture integration during comprehension. *Journal of Cognitive Neuroscience*, *23* (8), 1845-54.

Holler, Judith, Heather Shovelton, & Geoffrey Beattie (2009). Do iconic hand gestures really contribute to the communication of semantic information in a face-to-face context? *Journal of Nonverbal Behavior*, *33*, 73-88.

Kendon, Adam (1980). Gesticulation and speech: Two aspects of the process of utterance. In Mary R. Key (Ed.), *The relationship of verbal and nonverbal communication* (pp. 207–227). The Hague: Mouton and Co.

Kendon, Adam (1988). How gestures can become like words. In Fernando Poyatos (Ed.), *Cross-cultural perspectives in nonverbal communication* (pp. 131-41). Toronto: Hogrefe.

Kendon, Adam (2004). *Gesture: Visible action as utterance*. Cambridge, UK: CUP.

Kirchhof, Carolin (2011). So what's your affiliation with gesture? In Carolin Kirchhof, Zofia Malisz, & Petra Wagner (Eds.), *Proceedings of the 2nd conference on gesture and speech in interaction* (n.p.). Bielefeld: Digital copy.

Kopp, Stefan, & Ipke Wachsmuth (2004). Synthesizing multimodal utterances for conversational agents. *Journal of Computer Animation and Virtual Worlds, 15*, 39-52.

Krauss, Robert M., Yihsiu Chen, & Rebecca F. Gottesman (2000). Lexical gestures and lexical access: A process model. In David McNeill (Ed.), *Language and gesture* (pp. 261–283). Cambridge, UK: CUP.

Massaro, Dominic W., & Michael M. Cohen (1993). Perceiving asynchronous bimodal speech in

consonant-vowel and vowel syllables. *Speech Communication*, *13*, 127-134.

Massaro, Dominic W., Michael M. Cohen, & Paula M. T. Smeele (1996). Perception of asynchronous and conflicting visual and auditory speech. *Journal of the Acoustical Society of America*, *100* (3), 1777-1786.

McGurk, Harry, & John MacDonald (1976). Hearing lips and seeing voices. *Nature*, *264* (5588), 746–748.

McNeill, David (1985). So you think gestures are nonverbal? *Psychological Review*, *92* (3), 350-371.

McNeill, David (2005). *Gesture and thought*. Chicago, IL: University of Chicago Press.

McNeill, David (2012). *How language began: Gesture and speech in human evolution*. Cambridge, UK: CUP.

Nishida, Shin'ya (2006, July). *Interactions and integrations of multiple sensory channels in human brain*. Presented at the IEEE international conference on multimedia and expo, Toronto, ON.

Ojanen, Ville (2005). *Neurocognitive mechanisms of audiovisual perception*. Unpublished doctoral dissertation, Helsinki University of Technology.

Olshausen, Bruno A. (2000, October 10). *Aliasing*. Part of the seminar "PSC129 – Sensory Processes" held at Redwood Center for Theoretical Neuroscience, UCB, Berkeley.

Özyürek, Asli, Roel M. Willems, Sotaro Kita, & Peter Hagoort (2007). On-line integration of semantic information from speech and gesture: Insights from event-related brain potentials. *Journal of Cognitive Neuroscience*, *19* (4), 605-616.

Petrini, Katrin, Samuel P. Holt, & Frank Pollick (2010). Expertise with multisensory events eliminates the effect of biological motion rotation on audiovisual synchrony perception. *Journal of Vision*, *10* (5), 1-14.

Raithel, Jürgen (2006). *Quantitative Forschung*. Wiesbaden: VS Verlag für Sozialwissenschaften / GWV Fachverlage GmbH.

Schegloff, Emanuel A. (1985). On some gestures' relation to talk. In J. Maxwell Atkinson and John Heritage (Eds.), *Structures of social action. Studies in conversation analysis* (pp. 266–296). Cambridge: CUP.

Taylor, Humphrey (2007). The case for publishing (some) online polls. *The Polling Report*, *23* (1), n.p.

Thies, Alexandra, 2003. *First the hand, then the word: On gestural displacement in non-native English speech*. Unpublished SEII thesis, Bielefeld University.

Van Wassenhove, Virginie, Ken W. Grant, & David Poeppel (2002, April). *Temporal integration in the McGurk effect*. Poster presented at the 9th cognitive neuroscience annual meeting, San Francisco, CA.

Van Wassenhove, Virginie, Ken W. Grant, & David Poeppel (2007). Temporal window of integration in auditory–visual speech perception. *Neuropsychologia*, *45*, 598–607.

Vatakis, Argiro, Jordi Navarra, Salvador Soto-Faraco, & Charles Spence (2008). Audiovisual temporal adaptation of speech: temporal order versus simultaneity judgments. *Experimental Brain Research*, *185* (3), 521-529.

Wheatland, Nkenge, Yingying Wang, Huaguang Song, Michael Neff, Victor Zordan, & Sophie Jörg (2015). State of the art in hand and finger modeling and animation. *Computer Graphics Forum, 34*, 735-760.

Wilcox, Rand R. (2005). *Introduction to robust estimation and hypothesis testing (2nd ed.)*. Burlington, MA: Elsevier.

Wittenburg, Peter, Hennie Brugman, Albert Russel, Alex Klassmann, & Han Sloetjes (2006, May). *ELAN: A professional framework for multimodality research*. Presented at the 5th international conference on language resources and evaluation (LREC), Genoa, IT.